Voices in a Crowd: Searching for Clusters of Unique Perspectives

Nikolas Vitsakis†
Heriot-Watt University
nv2006@hw.ac.uk

Amit Parekh†
Heriot-Watt University
amit.parekh@hw.ac.uk

Ioannis Konstas
Heriot-Watt University
i.konstas@hw.ac.uk

Abstract

Language models have been shown to reproduce underlying biases existing in their training data, which is the majority perspective by default. Proposed solutions aim to capture minority perspectives by either modelling annotator disagreements or grouping annotators based on shared metadata, both of which face significant challenges. We propose a framework that trains models without encoding annotator metadata, extracts latent embeddings informed by annotator behaviour, and creates clusters of similar opinions, that we refer to as *voices*. Resulting clusters are validated post-hoc via internal and external quantitative metrics, as well a qualitative analysis to identify the type of voice that each cluster represents. Our results demonstrate the strong generalisation capability of our framework, indicated by resulting clusters being adequately robust, while also capturing minority perspectives based on different demographic factors throughout two distinct datasets.[1]

Content Warning: This document contains and discusses examples of potentially offensive and toxic language.

1 Introduction

Supervised training is rooted in the presupposition that every example in a dataset has a single ground truth, also known as the gold label (Hettiachchi et al., 2021). However, disagreement among dataset annotators challenges the notion that a single, per-example, ground truth exists (Uma et al., 2022a,b). While disagreement can be indicative of task difficulty or semantic ambiguity (Jiang and Marneffe, 2022; Sandri et al., 2023; Wang et al., 2021), it can also indicate the existence of both stable *and* conflicting inter-annotator perspectives (Abercrombie et al., 2023; Basile, 2020).

Nevertheless, capturing minority perspectives present in the data, which we parallel to *voices in a crowd*, has proven challenging. Two main approaches attempt to move beyond gold labels: **i) disagreement-based** which leverage annotator disagreement to provide distributional per-item prediction labels (Leonardelli et al., 2023; Uma et al., 2022a,b), and **ii) metadata-based**, which encode annotator metadata to boost the signal from voices with the same metadata labels (Beck et al., 2024; Fleisig et al., 2023; Gupta et al., 2023) (i and ii respectively in Figure 1).

However, both approaches come with strong vulnerabilities. **Disagreement-based** approaches collapse multiple minority voices into a singular, per-item, minority-majority distribution (Gordon et al., 2022), essentially limiting the number of expressed voices to the number of predicted labels (i.e., two voices in a binary prediction task). On the other hand, while **metadata-based** approaches allow for multiple minority voices to be expressed (albeit limited by metadata collected), they are based on the erroneous assumption that most members that share metadata labels (e.g., gendered females) will also exhibit similar patterns of behaviour (Beck et al., 2024; Dang et al., 2020).

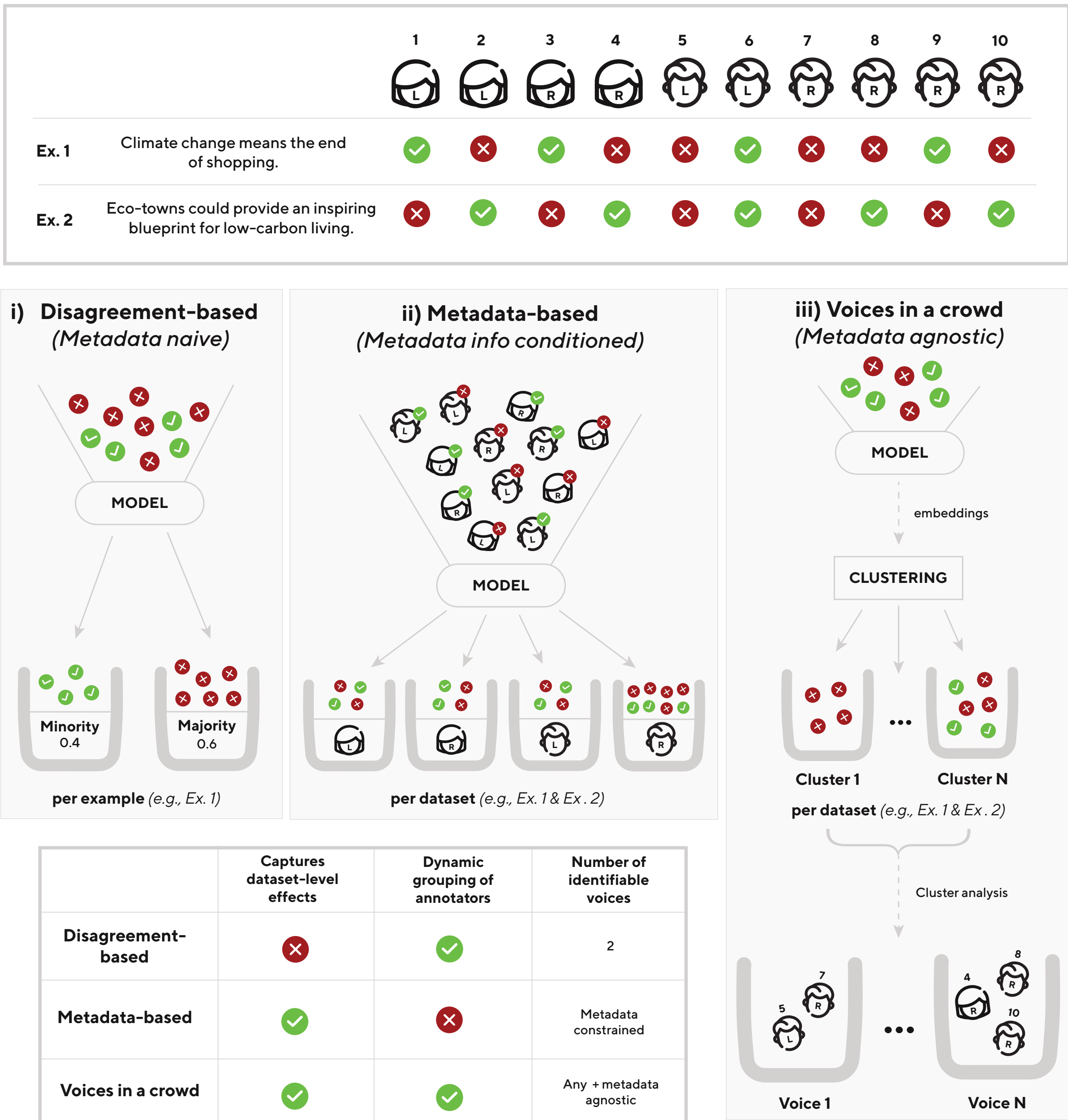
We introduce a framework that addresses both issues (Figure 1iii): it forms multiple clusters of distinct voices *solely* based on *annotator behaviours exhibited during the annotation process* in an unsupervised manner. Our pipeline trains models to predict *each annotation made by each annotator* for a given text input. The final hidden states form what we refer to as *behavioural embeddings*, representing how a given annotator will behave when shown that text sample. These are then clustered via unsupervised methods, with each cluster formed being a potential *voice*—a group perspective of annotators with similar annotating behaviours.

We apply our framework to two datasets related to political bias that have been found to contain multiple heterogeneous and conflicting perspectives (Chen et al., 2019; de Zarate et al., 2020; Menini and Tonelli, 2016; Németh, 2023). To identify the group whose voice each cluster belongs

†Equal contribution

[1]All code is made available at https://github.com/Ni-Vi/Cluster.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12517–12539
November 12-16, 2024 ©2024 Association for Computational Linguistics



	Captures dataset-level effects	Dynamic grouping of annotators	Number of identifiable voices
Disagreement-based	✗	✓	2
Metadata-based	✓	✗	Metadata constrained
Voices in a crowd	✓	✓	Any + metadata agnostic

Figure 1: Different approaches for handling annotations: i) disagreement-based create per-example distributional labels which fail to account for dataset-level effects; ii) metadata-based train models on annotations linked with annotator metadata, which often groups disagreeing annotators who share metadata labels; iii) the *"Voices in a crowd"* approach dynamically creates clusters based on annotation patterns and finally verifies each cluster as a voice based on post-hoc matched metadata labels.

to, we match each data point with annotator metadata post-hoc while we also conduct an in-depth qualitative analysis of the clusters themselves. The resulting clusters show high internal label consistency of either **i) dataset majority labels** (e.g., left-leaning in a left-leaning majority dataset), **ii) dataset minority labels** (e.g., right-leaning in a left-leaning majority dataset), but most importantly their intersection resulting in **iii) inter-minority labels** (e.g., right-leaning and highly educated, in a left-leaning, non-highly educated majority dataset). We are the first to dynamically identify voices of minority opinions within larger majority/minority groups, highlighting the significance of providing an intersectional understanding of annotators that goes beyond current grouping methodologies.

2 Related Work

Disagreement-Based Solutions As an alternative to gold labels, recent research has introduced the use of silver labels, i.e., distributional per-item labels that measure disagreement amongst annotators (Davani et al., 2022; Leonardelli et al., 2023; Uma et al., 2022a,b). While such approaches can allow for the identification of controversial examples in datasets (Fornaciari et al., 2022), they fail to capture stable inter-annotator disagreements throughout the dataset that could provide insight as to why disagreement occurs beyond an item-by-item scale (Abercrombie et al., 2023; Vitsakis et al., 2023).

To be more specific, disagreement-based solutions essentially limit the number of possible expressed voices into the number of predicted labels; the upper bound of possible voices expressed in a binary task is always two, no matter how diverse the dataset. Unfortunately, this type of aggregation leads to the erasure of what we define as *inter-minority voices*: stable opinions held by minority groups that are in conflict with each other as well as the majority, across examples. While there are some disagreement-based approaches that attempt a more nuanced expression of varied voices present in the data (Casola et al., 2023; Lo and Basile, 2023; Mokhberian et al., 2024), there is still a distinct lack of a holistic framework—such as ours—that both clusters and explains the type of voice expressed.

Metadata-Based Solutions A recent trend aiming to capture diverse perspectives has attempted to group annotators based on their metadata. Such approaches encode collected annotator metadata, such as annotator beliefs (Davani et al., 2023; Rottger et al., 2022) or demographics (Fleisig et al., 2023; Gupta et al., 2023), into the training pipeline to allow learning of patterns between annotations and in-group tendencies. While the incorporation of such information can seemingly improve model performance in specific tasks (Welch et al., 2020), evidence suggests that such results might be dataset-specific (Lee et al., 2023).

This is due to the assumption that annotators sharing metadata labels will behave similarly during the annotation process. However, demographics are not necessarily predictive of underlying behaviour (Beck et al., 2024; Hwang et al., 2023), while social sciences have also explained similar issues with self-reported measures (Dang et al., 2020; Schwarz, 1999). With the added issue that annotator metadata is often not collected outright (Prabhakaran et al., 2021), there is a direct need for methodologies that identify distinct group voices based on factors other than a-priori collected labels.

Unsupervised Learning and Clusters of Voices To circumvent previously mentioned issues, unsupervised learning could be employed along the lines of how past research identified emergent themes within corpora via clustering of latent textual embeddings (Dhillon and Modha, 2001; Meng et al., 2022; Sevillano et al., 2007; Wich et al., 2020). Recently, Meng et al. (2022) showed promising results in automatic topic discovery by utilising pretrained language models to cluster representations in a *joint latent space*: formed by combining latent spaces of multiple modalities during learning, in this case word and document level embeddings. We aim to take this work further through our use of joint behavioural embeddings, informed by both text and annotating behaviour, to automatically find *voices*, i.e., clusters of similar opinions.

There are significant challenges to this approach. Fine-tuning pretrained language model embeddings produce embeddings that are often anisotropic and anisometric (Rajaee and Pilehvar, 2021; Xu and Koehn, 2021); which paired with their high dimensionality nature, makes clustering via distance-based metrics challenging. However, by using appropriate dimensionality reductions (Cai et al., 2020; Mu and Viswanath, 2018), the relationships between features can be analysed and clustered through Euclidean distance-based metrics (McInnes et al., 2020).

3 Experimental Setup

Our framework comprises a **supervised** and an **unsupervised component**. The former produces latent embeddings informed by both text and annotating behaviour that the latter uses to cluster into voices. Being the first such approach, we compared performance across a variety of transformer-based architectures, clustering, and dimensionality reduction techniques to identify optimal combinations.

The **supervised component** explores several modelling choices (Section 4) fine-tuned on each dataset to predict each annotator's individual annotation for a given example without providing any annotator metadata that could bias the model (Vitsakis et al., 2023). The **unsupervised component** then performs dimensionality reduction on the *behavioural embeddings*—the final hidden states from the supervised component—and finally creates clusters via several unsupervised algorithms (Section 5). Clusters are evaluated through internal (i.e., intra-cluster similarity) and external metrics (i.e., consistency of demographic labels in a given cluster), and via qualitative analysis of the best-performing combination of components.

3.1 Datasets

All datasets used in our experiments contain the following annotator demographics: personal political leaning, and education level. We explicitly chose datasets that are political in nature, as self-declared political affiliation metadata should largely match with similarities in annotating behaviour.

Media Bias Annotation Dataset (MBIC; Spinde et al., 2021a,b) comprises sentences from media articles that may contain political bias from news outlets across the political spectrum (e.g., Fox News, MSNBC, etc.) covering 14 potentially divisive topics (e.g., gender issues, coronavirus, the 2020 American election). 784 crowd-sourced annotators labelled sentences on whether they consider them to contain bias. Demographics were slightly skewed in political ideology (44.3% left-leaning, 26.7% right-leaning, 29.1% center).

Global Warming Stance Dataset (GWSD; Luo et al., 2020) contains opinions of varying intensities on the subject of global warming, gathered from news outlets with different political leanings (e.g., The New York Times, Breitbart). 398 annotators labelled each sentence with whether they agreed, disagreed, or were neutral. Demographic skew of this dataset mirrored that of MBIC in self-reported political affiliation (46% Democrat, 21.2% Republican, 28.8% Independent, 4% Other).

4 Supervised Component

Each of the following modelling architectures was trained through a different combination of inputs (visual representation in Appendix B): given a text sample in a dataset, $\mathbf{x} \in \mathbf{X}$, we predict the *individual annotation* of each annotator $p_\theta(\mathbf{y}|\mathbf{x})$ where $\mathbf{y} = (y_1, \ldots, y_K)$ and K is the total number of unique annotators within the dataset.

Unpooled Cross Attention uses a pretrained T5 encoder (Raffel et al., 2020) where the encoded text and the embedded annotator unique identifiers are fed through a decoder to predict each annotator's annotation as a sequence. Annotator embeddings are directly informed by the text via a cross-attention layer aiming to capture the influence of the text in the annotators' behaviours.

Pooled Cross Attention follows Sullivan et al. (2023), which showed strong performance in predicting annotator disagreement in the 2023 Learning With Disagreements shared task (LeWiDi; Leonardelli et al., 2023). This model is similar in structure to *Unpooled Cross Attention* since it also uses a T5 encoder as the backbone. However, the dimension for each encoded text token is downsampled, as previous research has indicated possible benefits in the salience of encoded features (Dhingra et al., 2018; Holzenberger et al., 2018; Schick and Schütze, 2019). Finally, decoder outputs are pooled (Reimers and Gurevych, 2019) to predict an aggregated annotation for each batch.

Encoder-Encoder treats text and annotators as separate modalities, inspired by multimodal approaches (Agarwal et al., 2020; Singh et al., 2022; Tan and Bansal, 2019). The encoded text (using T5) and embedded annotator IDs are concatenated and fed through a bidirectional encoder to predict the annotation of each annotator, allowing for interaction between text and annotator embeddings.

Classifier Model simply concatenates the text with the unique annotator identifier, before passing to an encoder (BERT; Devlin et al., 2019 for GWSD, and RoBERTa; Liu et al., 2019 for MBIC) to predict each annotation label independently. The independence between annotators limits interaction between annotators during training.

Pretrained Decoder is a decoder-only GPT-2 model (Radford et al., 2019) prompted with the concatenated text and annotator identifiers in the form "`<text> [SEP] <Ann 1> [SEP] ... <Ann K>`" and predicts the annotation for each annotator.

Pretrained Encoder-Decoder similarly to *Unpooled Cross Attention*. It uses a pretrained T5 encoder-decoder where unique annotator identifiers are embedded through the decoder—instead of a decoder trained from scratch—to predict each annotator's annotation autoregressively. Since the decoder is unidirectional, it forces causal attention across annotators in their canonical order.

Metrics We compute F1 score to measure the accuracy of predictions, and Average Pairwise Cosine Similarity (APCS) between hidden states of predicted annotations to illustrate how dense the latent states are by the end of training; we show that lower scores generally correlate with better clustering performance (see Section 5.1).

	F1 Score ↑	APCS ↓
GWSD Dataset		
Unpooled Cross Attention	**0.65**	**0.14 ± 0.07**
Pooled Cross Attention	0.19	0.54 ± 0.13
Encoder-Encoder	0.63	0.15 ± 0.11
Classifier Model	0.63	0.81 ± 0.14
Pretrained Decoder	0.62	0.66 ± 0.08
Pretrained Encoder-Decoder	0.19	0.95 ± 0.02
MBIC Dataset		
Unpooled Cross Attention	**0.72**	0.22 ± 0.05
Pooled Cross Attention	0.43	0.70 ± 0.06
Encoder-Encoder	**0.72**	**0.21 ± 0.06**
Classifier	0.38	1.00
Pretrained Decoder	0.63	0.75 ± 0.07
Pretrained Encoder-Decoder	0.71	0.74 ± 0.25

Table 1: Overall performance (F1 Score) for the supervised component of our framework (6 modelling architectures) on MBIC and GWSD for the task of individual annotator prediction. We also report the Average Pairwise Cosine Similarity (APCS) across the final hidden states; lower scores indicate greater variety in representation correlating with better clustering performance.

Results Table 1 summarises the results. For GWSD, Unpooled Cross Attention achieved the highest F1 score and lowest APCS, whereas it shared a similar performance with Encoder-Encoder for MBIC (albeit the latter has slightly lower APCS). This could be down to the bidirectional attention mechanism (either through cross-attention or encoder self-attention) between the annotator embeddings and the text during training.

These results also showcase the importance of reporting on the quality of the hidden states. For example, while the Pretrained Encoder-Decoder and Classifier Model have high F1 scores on the MBIC and GWSD datasets respectively, their low scores on APCS indicate dense hidden states that would result in poor clustering outcomes. Overall, our findings show that the bidirectional attention-based models that allow interaction between text and annotator embeddings are the only consistent architectures to show high F1 and low APCS scores.

5 Unsupervised Component

Dimensionality Reduction We perform dimensionality reduction on the hidden states before clustering as follows: a baseline without dimensionality reduction, Principal Component Analysis (PCA; a linear combination of components) and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP; a non-linear transformation algorithm; McInnes et al., 2020). Both PCA (Gupta et al., 2020; Sia et al., 2020) and UMAP (Ait-Saada and Nadif, 2023; Cai et al., 2020; George and Sumathy, 2023) improve feature representation in high-dimensional latent spaces leading to improved clustering.

Clustering Algorithms We used three clustering techniques: K-means (MacQueen et al., 1967; Pedregosa et al., 2011), Gaussian Mixture Models (GMM; Rasmussen, 1999), and HDBSCAN (McInnes et al., 2017). Each of these techniques have been used to cluster features when paired with either PCA (Asyaky and Mandala, 2021; Hosseini and Varzaneh, 2022; Liu et al., 2021), or UMAP (Allaoui et al., 2020; Asyaky and Mandala, 2021).

Metrics We use two **internal validation** metrics to assess average similarity scores between clusters, namely *Silhouette* (Pedregosa et al., 2011; Rousseeuw, 1987) and *Davies-Bouldin Index* (Davies and Bouldin, 1979; Pedregosa et al., 2011). Silhouette assesses intra-cluster separation and is bound between -1 and 1, with 1 being the best possible score, with a threshold of 0.5 for moderate clusters (Lengyel and Botta-Dukát, 2019; Shahapure and Nicholas, 2020). The Davies-Bouldin Index measures intra-cluster dissimilarity, with 0 indicating the lowest possible score (Idrus, 2022; Kärkkäinen and Fränti, 2000).

We use *Purity* to assess the **external validity** of clusters. Purity measures the internal consistency of assigned labels within a cluster and evaluates whether a cluster is prototypical (i.e., representative) across provided labels within a dataset (Christodoulopoulos et al., 2010). In our case, we report both average purity and the percentage of prototypical clusters per method. We define a cluster as prototypical if its metadata label purity (i.e., political leaning and education level) is significantly different from the original dataset metadata label distribution with a threshold of ± 10%. These metrics allow us to automatically assess whether a cluster emerging from annotator behaviours during training is linked to any of the annotator labels (e.g., a cluster with high right-leaning metadata label purity) and thus is indicative of a distinct voice.

5.1 Results

Optimal cluster numbers were automatically calculated using hyperparameter sweeps to maximise

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑		*Prototypical Cluster %* ↑	
				Political	Education	Political	Education
Unpooled Cross Attention							
No dim. reduction	19	6.35	0.02	**0.71**	**0.71**	15.8	0.0
w/ PCA	10	1.98	0.10	0.36	0.43	20.0	0.0
w/ UMAP	19	0.81	0.47	0.38	0.42	31.6	5.3
Pooled Cross Attention							
No dim. reduction	19	3.03	0.06	0.42	0.48	26.0	5.3
w/ PCA	19	1.04	0.28	0.47	0.46	5.5	0.0
w/ UMAP	12	1.13	0.29	0.70	0.50	25.0	8.0
Encoder-Encoder							
No dim. reduction	19	6.93	0.01	0.41	0.46	21.1	15.8
w/ PCA	19	0.49	**0.54**	0.53	0.43	15.0	7.7
w/ UMAP	19	**0.49**	0.53	0.51	0.48	**36.8**	**21.1**
Classifier Model							
No dim. reduction	5	1.98	0.06	0.49	0.44	0.0	0.0
w/ PCA	13	0.84	0.36	0.44	0.44	7.4	0.0
w/ UMAP	18	0.55	0.49	0.44	0.49	5.5	5.5
Pretrained Decoder							
No dim. reduction	19	2.76	0.06	0.39	0.42	16.0	11.1
w/ PCA	18	1.89	0.12	0.44	0.61	5.6	5.6
w/ UMAP	19	1.01	0.34	0.36	0.42	11.0	11.0
Pretrained Encoder-Decoder							
No dim. reduction	5	1.62	0.16	0.44	0.48	0.0	0.0
w/ PCA	8	1.74	0.20	0.37	0.46	0.0	0.0
w/ UMAP	5	0.75	0.44	0.46	0.46	0.0	0.0

Table 2: Internal and external validation metrics for the unsupervised component with the **K-Means clustering algorithm on the MBIC dataset**. Internal validation metrics explain intra-cluster separation through higher Silhouette and lower Davies-Bouldin (DB Index) scores. External validity indicates the potential capturing of a voice, measured by the average Purity score and % of prototypical clusters.

the Silhouette score (see Appendix A for more information). Table 2 shows the clustering of our best performance combination, K-means with a UMAP dimensionality reduction on the MBIC dataset as other configurations performed less optimally as seen in Appendix C.

Internal Validity Metrics Dimensionality reduction significantly impacted the quality of the resulting clusters; UMAP outperformed PCA, while no dimensionality reduction showed the worst overall results (for averages, see Appendix A.2). The only exception was Encoder-Encoder, where PCA and UMAP perform comparably.

Encoder-Encoder performed best overall: being the only model with Silhouette and Davies-Boulding Index scores above/below the respective cutoff points of 0.5, indicating adequate intra-cluster separation for both metrics (Idrus, 2022; Lengyel and Botta-Dukát, 2019; Shahapure and Nicholas, 2020). Interestingly, the Classifier Model also performed relatively well despite being the lowest-performing of the supervised component.

External Validity Metrics Average purity scores are largely inconclusive as higher scores are not always linked with better performance, which is evident through comparisons with other evaluative metrics. For example, Unpooled Cross Attention with no dimensionality reduction, scores poorly on internal validation metrics, while average purity is the highest across both metadata labels.

Overall, these findings echo those seen in Table 1, where models with the lowest APCS scores also had the best performance in internal and external validation metrics. The best-performing model was Encoder-Encoder with UMAP outperforming PCA, followed by Unpooled Cross Attention. While UMAP only marginally outperformed PCA in terms of internal validation scores, the la-

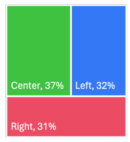
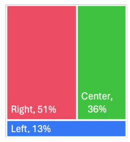
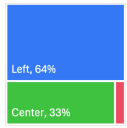
Dataset/Cluster No.	Examples	Bias Label	Distribution
MBIC -1	British Olympic swimmer Sharron Davies also slammed the concept of transgender athletes.	✓	
	BBC Presenter Gabby Logan has said that it is not fair that transgender women can compete in sport alongside biologically female women.	✓	
	BBC Presenter Gabby Logan has said that it is not fair that transgender women can compete in sport alongside biologically female women.	✗	
MBIC -7	Trump — who has been criticized for painting an overly rosy picture of the outbreak, often contradicting his own health officials - insisted on Friday that his administration was “magnificently organized” and “totally prepared" to address the virus.	✓	
	Google declined to offer details beyond Huntley’s tweets, but the unusually public attribution is a sign of how sensitive Americans have become to digital espionage efforts aimed at political campaigns.	✗	
	At least 25 transgender or gender-nonconforming people were killed in violent attacks in the United States last year, according to the Human Rights Campaign, which has been tracking anti-trans violence since at least 2015.	✓	
MBIC -8	Though conservatives try to demonize Ocasio-Cortez an Omar, their actual policy views are perfectly mainstream. The New York lawmaker proposed a 70 percent tax on top incomes — a view backed by public opinion and many well-respected economists.	✗	
	British Olympic swimmer Sharron Davies also slammed the concept of transgender athletes.	✗	
	At least 25 transgender or gender-nonconforming people were killed in violent attacks in the United States last year, according to the Human Rights Campaign, which has been tracking anti-trans violence since at least 2015.	✓	

Table 3: Analysis of clusters on the MBIC dataset with the Encoder-Encoder architecture and UMAP dimensionality reduction. We report the cluster number, representative examples of the cluster, and their paired annotation (✓ for perceived bias, ✗for no perceived bias). We also show the distribution of annotator characteristics which is indicative of the prototypical nature of each cluster.

bel distributions in the clusters resulting from PCA were minimally different when compared to label distributions present in the original data. Finally, we found that prototypical clustering percentage was a strong indicator of capturing representative clusters of voices.

Manual inspection of PCA-formed clusters indicated that clusters formation was mostly based around the most salient features discovered during training, namely the unique annotator tokens or the inter-sentence similarities. A possible reason for this phenomenon could be that PCA reduces dimensionalities to the most salient principal components, which are not conducive to clustering based on contextual features in large language models (Cai et al., 2020). Interestingly, this phenomenon was reproduced with UMAP when instructing the model to focus on finding clusters based on local and not overarching features (McInnes et al., 2020).[2]

6 Qualitative case study

While encouraging, our findings cannot be simply explained through either internal or external validation metrics. To assess whether a cluster is truly indicative of a voice, we looked at the content of the clusters themselves. High purity of a cluster should be reflected in the text-annotation pair content (e.g., a cluster with high left-leaning purity should be paired with left-leaning opinions).

Furthermore, this relationship between labels and opinions contained in each voice should also mediated by representation of other metadata labels: clusters predominantly represented by a single label should denote opinions held by that group and not others, while an increasing representation of other group labels should denote opinions that are less divisive between groups. For example, a cluster with a high concentration of center and left-leaning metadata labels, but not right-leaning ones, should contain opinions that are less divisive between the former, but divisive with the latter.

Given our labels, this can result in three distinct types of voices: majority, minority and inter-minority. Majority voice clusters consist of high purity of a majority metadata label (e.g., left-leaning opinions in a left-leaning majority dataset), while minority voices are the same for dataset minority labels (e.g., right-leaning opinions in a left-leaning majority dataset), and inter-minority voices, which are clusters that consist of high purity across combination of metadata labels (e.g., high purity in both right-leaning and highly educated metadata labels in a dataset with left-leaning and non-highly educated majority metadata labels).

[2]A possible solution to this issue is to remove the top principal components resulting in more salient representations, and thus improve clustering performance (Mu and Viswanath, 2018); we leave this for future work.

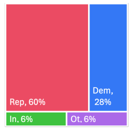
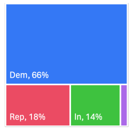
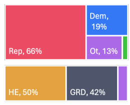
Dataset/Cluster No.	Examples	Agreement Label	Distribution
GWSD -9	The early 21st-century drought that afflicted Central Asia is the worst in Mongolia in more than 1,000 years, and made harsher by the higher temperatures consistent with man-made global warming.	✓	
	Climate change means the end of shopping.	∼	
	The oil sands are responsible for just 0.001 percent of global greenhouse emissions	∼	
GWSD -2	There is a connection between human activity and an assumptive change in global climate.	✓	
	Hiring a White House "climate change czar" would be a good idea.	✓	
	Scaring young people into believing that climate change is going to kill young people is child abuse.	✗	
GWSD -5	The oil sands are responsible for just 0.001 percent of global greenhouse emissions	✓	
	This could mean that current I.P.C.C. model predictions for the next century are wrong, and there will be no cooling in the North Atlantic to partially offset the effects of global climate change over North America and Europe.	✓	
	Eco-towns could provide an inspiring blueprint for low-carbon living	✗	

Table 4: Analysis of clusters on the GWSD dataset using same parameters as the MBIC dataset, and results are shown in a similar fashion (✓agree with the statement, ✗for disagree and ∼ for neutral). Distribution of annotator characteristics is provided.

To extract our clusters, we used the best-performing combination, i.e., Encoder-Encoder with UMAP and K-means clustering. As the purpose of the case study was to show examples of what can be achieved through our framework, we chose examples of prototypical clusters, indicative of the variety of voices in each dataset, and found multiple examples of each effect described. We pick three prototypical clusters from a single clustering run, each representing a distinct voice, and discuss them in Table 3 and Table 4.

6.1 MBIC Dataset

MBIC-1: Minority Voice This cluster is a prototypical example of minority-led consensus amongst annotators. The cluster's distribution is more even, following the original label distribution closer (44.3%, 29.1%, 26.7% for left, center, and right political lean). Such clusters often contain different annotations for the same sentences, while there is no strong emerging effect from collected labels.

MBIC-7: Minority Voice This is a minority voice, with the distribution of labels indicating that the cluster is primarily formed of right-leaning opinions. While Item 1 is expectantly labelled as 'bias', Item 3 contains no obvious biased words, despite coming from an obvious place of concern for a marginalised minority.

MBIC-8: Majority Voice This is an example of a majority dominant cluster. Such clusters are populated by the opinion of the original dataset's distributional majority label although with a much heavier skew, indicating a stable and consistent behaviour of the group. The labelling distribution of this cluster is expected to be populated by left-leaning views and indeed sentences that were previously labelled as biased in non-left-leaning clusters (Item 1 of Cluster 1, and Item 3 of Cluster 7), were consistently found not to be labelled as such.

6.2 GWSD Dataset

GWSD-9: Minority Voice This is an example of a minority cluster, as indicated by the differences in the distribution of the minority label between the cluster and the original data (21% in the original data, 60% representation in this cluster). While the expressed opinions within were generally agreeable about climate-changing effects, there was no agreement with more politically charged statements.

GWSD-2: Majority Voice This is a majority-dominant cluster. Opinions that could be perceived as more explicitly political were found to be increasingly common (Item 2), while there was also evidence of general agreement with some strongly politically charged examples (Item 3).

GWSD-5: Minority-Minority Voice An example of a minority within a minority perspective. Opinions are over-represented by two minority labels, the "republican" in terms of political affiliation, and that of the "higher degree" in terms of education level (8.4% label representation in the original dataset). Opinions showed fewer "neutral" responses and were generally indicative of a well-informed audience, explicitly agreeing with more technical items such as Item 2 and especially Item 1, which received mostly “neutral” scores in other clusters (e.g., Cluster 9).

7 Conclusion

We propose a novel framework to identify underlying minority perspectives in data. We compared six distinct model architectures trained on a classification task, without providing any annotator metadata to avoid biasing their training. Subsequently, final hidden states were passed through various methods of dimensionality reduction (UMAP and PCA), with the resulting embeddings used to create clusters through various unsupervised algorithms (K-means, GMM, and HDBSCAN).

The resulting clusters were adequately separated according to internal and external validation metrics. Further qualitative analysis of clusters produced by our best-performing model showcased the ability of our framework to capture perspectives as shown by three distinct types: clusters representative of a minority, a majority, and clusters that captured multiple minority labels, i.e., a minority within a minority.

Limitations & Ethical Considerations

Internal & External Validity Related As shown in Table 2 and Appendix C while internal validation scores *can* be indicative of well-defined clusters of minority perspectives, they are not necessarily so. We further explain in Appendix C that this might be due to our training on unique annotator tokens, which could hinder organic clustering based on behaviour by providing an alternative and easier to learn signal in unique annotator tokens. Finally, for comparisons between preliminary results between different model sizes alongside a brief discussion on their impact, see Appendix C.6.

We aim to expand upon this in future work, by modifying training of our supervised component to incorporate aspects more representative of group behaviours such as inter and intra annotator disagreement (Abercrombie et al., 2023; Casola et al., 2023; Lo and Basile, 2023; Mokhberian et al., 2024; Uma et al., 2021). This would assess the limitations of disagreement-based approaches described in Section 2 by enhancing group behavioural signals, as indicated by annotator agreement/disagreement (Deng et al., 2023; Mokhberian et al., 2024), while also allowing for such effects to be captured on the dataset-level. Furthermore, incorporation of such methodologies into our framework would further address the limitations of disagreement-based methodologies by allowing for any number of voices to be expressed.

Automatic Detection of Voices A current limitation of the framework is the ability to automatically assess the performance of each combination without manual inspection. While necessary at this step to prove the efficacy of our framework, we aim to expand this in future work by introducing a a component that automatically extracts information from each cluster to allow for identification of voice without the need of matching clusters with metadata labels post-hoc.

We aim to employ a similar methodology to Fleisig et al. (2023), whose pipeline includes a GPT-2 based component that predicts the demographic group targeted by a given text. We aim to include similar components to extrapolate attitudinal and behavioural indicators of formed clusters via analysing the text-annotation pairs to generate labels representative of each captured voice similarly to how research in sentiment analysis, has previously classified opinions on politically charged data (Ansari et al., 2020; Dorle and Pise, 2018; Kazienko et al., 2023).

Labels and further marginalisation of minorities Our model uses labels procured during data gathering to validate emergent clusters. However, the labelling gathering process can potentially be an erasing process towards minorities in and of itself (Chandrabose et al., 2021; Hovy and Prabhumoye, 2021). For example, the labelling process can discriminate against socially marginalised minorities by not providing options consistent with an individual's identity (Chandrabose et al., 2021; Jo and Gebru, 2020).

In our case, we encountered this limitation with the GWSD dataset (Luo et al., 2020), which collected categorical labels about political affiliation of participants. Beyond the three primary labels ("Democrat", "Independent", "Republican"), the rest were aggregated into the "other" label. This resulted in a minority so small that our clustering methodology could not adequately disentangle it from the rest. Future research should look into directions alongside those explained in Section 7, which also should address these concerns for future iterations of our framework.

Dual Use of the Model An unfortunate outcome of methodologies aim to capture and expressed more nuanced perspectives can lead to identification of marginalised minority perspectives in datasets, which can lead to concerning practice of their removal in order to enhance a model's general performance (Sun et al., 2019; Xu et al., 2021). Nevertheless, Gaci et al. (2023) has also proposed that methodologies that identify minority perspectives can be used to curate datasets in order to amplify voices of specific marginalised groups.

We urge researchers to be transparent in their intended use of our framework, and to follow ethical frameworks and solutions that have been previously highlighted by the field in from the data collection process to model training and intended use (Blodgett et al., 2020; Hovy and Prabhumoye, 2021; Leidner and Plachouras, 2017; Navigli et al., 2023; Shmueli et al., 2021).

Acknowledgements

We would first like to thank the reviewers for their extensive feedback and continued engagement throughout the discussion period. We also thank the various members of the iLab, and more specifically Georgios Pantazopoulos and Malvina Nikandrou with whom conversations and brainstorming made this work possible. We would also like to extend our thanks to Verena Rieser, whose guidance early-on was instrumental. This work was supported by the Heriot-Watt University high-performance computing facility (DMOG) and associated support services.

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement. *Preprint*, arXiv:2301.10684.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for Visual Dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Mira Ait-Saada and Mohamed Nadif. 2023. Is Anisotropy Truly Harmful? A Case Study on Text Clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203, Toronto, Canada. Association for Computational Linguistics.

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. 2020. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In *Image and Signal Processing*, pages 317–325, Cham. Springer International Publishing.

Mohd Zeeshan Ansari, Mohd-Bilal Aziz, MO Siddiqui, H Mehra, and KP Singh. 2020. Analysis of political sentiment orientations on twitter. *Procedia Computer Science*, 167:1821–1828.

Muhammad Sidik Asyaky and Rila Mandala. 2021. Improving the performance of hdbscan on short text clustering by using word embedding and umap. In *2021 8th international conference on advanced informatics: Concepts, theory and applications (ICAICTA)*, pages 1–6. IEEE.

Valerio Basile. 2020. It's the End of the Gold Standard as We Know It: Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks. In *AIxIA 2020 – Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, pages 441–453, Berlin, Heidelberg. Springer-Verlag.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2020. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*.

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Aravindan Chandrabose, Bharathi Raja Chakravarthi, et al. 2021. An overview of fairness in data–illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584.

Junhua Dang, Kevin M King, and Michael Inzlicht. 2020. Why are self-report and behavioral measures weakly correlated? *Trends in cognitive sciences*, 24(4):267–269.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Juan Manuel Ortiz de Zarate, Marco Di Giovanni, Esteban Zindel Feuerstein, and Marco Brambilla. 2020. Measuring controversy in social networks through nlp. In *International Symposium on String Processing and Information Retrieval*, pages 194–209. Springer.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Inderjit S Dhillon and Dharmendra S Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42:143–175.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. 2018. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Saurabh Dorle and Nitin Pise. 2018. Political sentiment analysis through social media. In *2018 second international conference on computing methodologies and communication (ICCMC)*, pages 869–873. IEEE.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

Tommaso Fornaciari, Alexandra Uma, Massimo Poesio, and Dirk Hovy. 2022. Hard and soft evaluation of NLP models with BOOtSTrap SAmpling - BooStSa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–134, Dublin, Ireland. Association for Computational Linguistics.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2023. Targeting the source: Selective data curation for debiasing nlp models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 276–294. Springer.

Lijimol George and P Sumathy. 2023. An integrated clustering and bert framework for improved topic modeling. *International Journal of Information Technology*, pages 1–9.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Soumyajit Gupta, Sooyong Lee, Maria De-Arteaga, and Matthew Lease. 2023. Same same, but different: Conditional multi-task learning for demographic-specific toxicity detection. In *Proceedings of the ACM Web Conference 2023*, pages 3689–3700.

Vivek Gupta, Ankit Kumar, Pegah Nokhiz, Harshit Gupta, and Partha Talukdar. 2020. Improving Document Classification with Multi-Sense Embeddings. In *ECAI 2020*, pages 2030–2037. IOS Press.

Danula Hettiachchi, Mike Schaekermann, Tristan J McKinney, and Matthew Lease. 2021. The challenge of variable effort crowdsourcing and how visible gold can help. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–26.

Nils Holzenberger, Mingxing Du, Julien Karadayi, Rachid Riad, and Emmanuel Dupoux. 2018. Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments. In *Interspeech 2018*. ISCA.

Soodeh Hosseini and Zahra Asghari Varzaneh. 2022. Deep text clustering using stacked autoencoder. *Multimedia Tools and Applications*, 81(8):10861–10881.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Ali Idrus. 2022. Distance analysis measuring for clustering using k-means and davies bouldin index algorithm. *TEM Journal*, 11(4):1871–1876.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Ismo Kärkkäinen and Pasi Fränti. 2000. Minimization of the value of davies-bouldin index. In *Proceedings of the IASTED International Conference on Signal Processing and Communications (SPC'2000). IASTED/ACTA Press*, pages 426–432.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Attila Lengyel and Zoltán Botta-Dukát. 2019. Silhouette width using generalized mean—a flexible method for assessing clustering efficiency. *Ecology and evolution*, 9(23):13231–13243.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Ziquan Liu, Lei Yu, Janet H Hsiao, and Antoni B Chan. 2021. Primal-gmm: Parametric manifold learning of gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3197–3211.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *NLPerspectives@ ECAI*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Preprint*, arXiv:1802.03426.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, pages 3143–3152.

Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2461–2470.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of* the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3042–3049, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carl Rasmussen. 1999. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Norbert Schwarz. 1999. Self-reports: How the questions shape the answers. *American psychologist*, 54(2):93.

Xavier Sevillano, Germán Cobo, Francesc Alías, and Joan Claudi Socoró. 2007. Text clustering on latent thematic spaces: Variants, strengths and weaknesses. In *International Conference on Independent Component Analysis and Signal Separation*, pages 794–801. Springer.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, and K. Donnay. 2021a. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. *Preprint*, arXiv:2105.11910.

Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021b. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505.

Michael Sullivan, Mohammed Yasin, and Cassandra L Jacobs. 2023. University at buffalo at semeval-2023 task 11: Masda–modelling annotator sensibilities through disaggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022a. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5:818451.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022b. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. iLab at SemEval-2023 Task 11 Le-Wi-Di: Modelling Disagreement or Modelling Perspectives? In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669, Toronto, Canada. Association for Computational Linguistics.

Dongsheng Wang, Prayag Tiwari, Mohammad Shorfuzzaman, and Ingo Schmitt. 2021. Deep neural learning on weighted datasets utilizing label disagreement from crowdsourcing. *Computer Networks*, 196:108227.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Haoran Xu and Philipp Koehn. 2021. Cross-Lingual BERT Contextual Embedding Space Mapping with Isotropic and Isometric Conditions. *Preprint*, arXiv:2107.09186.

A Training Details

To aid in reproducibility, we report all training details and any relevant hyperparameters.

A.1 Hyperparameters

All models were trained using a single NVIDIA A40 GPU. A total of 1080 hours were used during training of all models. For all models, we used the AdamW optimizer (Loshchilov and Hutter, 2019) during training with weight decay 0.01. We report hyperparameters for each model and dataset in Table 5.From small performance gains during preliminary experiments, we disable bias across all linear layers.

Cluster training hyperparameters can be found in Table 6. Across every model, we found that when comparing hyperparameters for both PCA and UMAP converged to the same choices. For both methods, we found that 2 components yielded the best results. Additionally, for UMAP, we found that the optimal number of neighbours were found to be between 80–100 across all models,with a minimum distance ranging from 0.8 to 1 to yield better clustering performance.

A.2 Dimensionality Reduction

We report internal validity evaluation score averages across dimensionality reduction techniques in Table 7.

B Visual Representation of Models used in Training Component

Visual depictions of all model architectures seen in Figure 2.

C Cluster Metrics

C.1 GWSD Cluster Validity Scores - Kmeans

We report the GWSD internal and external validation metrics resulting from our clustering using a k-means algorithm and our various employed dimensionality reduction techniques in Table 8.

C.2 GWSD Cluster Validity Scores - GMM

We report the GWSD internal and external validation metrics resulting from our clustering using a GMM algorithm and our various employed dimensionality reduction techniques in Table 9. This methodology resulted in cluster metrics which were not as optimal as those of the K-means solutions.

Hyperparameter	Value
Unpooled Cross Attention	
Model name	`google/t5-v1_1-large`
Downsampling n. of layers	0-3
N. warmup steps	0- 800
Learning rate	0.0001 - 1e-08
Pooled Cross Attention	
Model name	`google/t5-v1_1-large`
Ann dim. factor	1-6
Downsampling n. of layers	0-3
N. warmup steps	0- 800
Learning rate	0.0001 - 1e-08
Encoder-Encoder	
Model name	`google/t5-v1_1-large`
Downsampling n. of layers	0-3
N. warmup steps	0- 800
Learning rate	0.0001 - 1e-08
Classifier Model	
Model name	`roberta-large`
N. warmup steps	0- 800
Learning rate	1e-11 - 1e-3
Pretrained Decoder	
Model name	`gpt2-large`
Downsampling n. of layers	0-3
N. warmup steps	0- 800
Learning rate	0.0001 - 1e-08
Pretrained Encoder-Decoder	
Model name	`google/t5-v1_1-large`
Downsampling n. of layers	0-3
N. warmup steps	0- 800
Learning rate	0.0001 - 1e-08

Table 5: Hyperparameters for all supervised models on each of our chosen datasets, obtained from running a hyperparameter sweep for 12 hours.

Hyperparameter	Value
PCA	
Cluster ranges	2 - 19
N components	2-40
GMM	
Cluster ranges	2-19
HDBSCAN	
Eps	0.0 - 1.0
Min samples	2 - 100
Min cluster size	2 - 100

Table 6: Hyperparameters for all clustering methods on each of our chosen datasets, obtained from running a hyperparameter sweep for 12 hours.

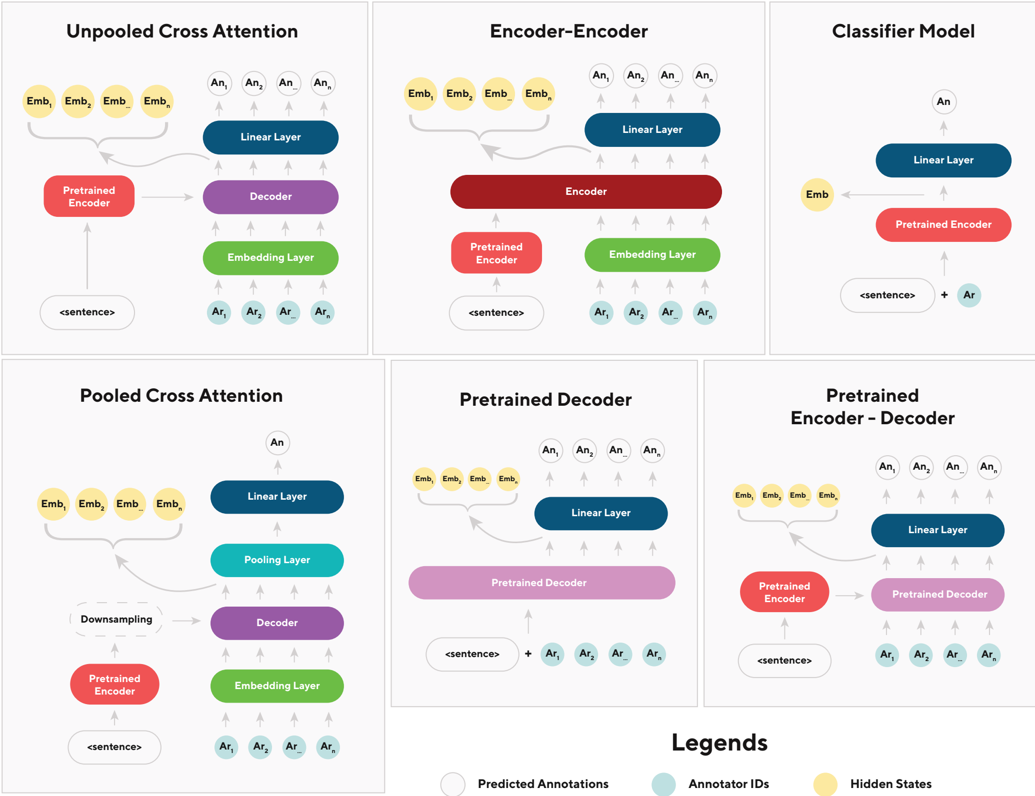


Figure 2: Training component: 6 modelling architectures for extracting hidden states (denoted with a yellow circle as Emb_n) used as input for the Clustering component.

	Davies-Bouldin Index	Silhouette
No dim. reduction	3.655	0.073
w/ PCA	0.491	0.56
w/ UMAP	0.565	0.53

Table 7: Dimensionality reduction effect on internal validity scores

C.3 GWSD Cluster Validity Scores - HDBSCAN

We report the GWSD internal and external validation metrics resulting from our clustering using an HDBSCAN algorithm and our various employed dimensionality reduction techniques in Table 10. Unfortunately, this methodology resulted in either large cluster numbers too large to be adequately analysed manually, or with metrics not as optimal as those of the K-means solutions.

C.4 MBIC Cluster Validity Scores- GMM

We report the MBIC internal and external validation metrics resulting from our clustering using a GMM algorithm and our various employed dimensionality reduction techniques in Table 11. Unfortunately, this methodology also resulted in cluster metrics which were not as optimal as those of the K-means solutions.

C.5 MBIC Cluster Validity Scores- HDBSCAN

We report the MBIC internal and external validation metrics resulting from our clustering using a HDBSCAN algorithm and our various employed dimensionality reduction techniques in Table 12. Unfortunately, this methodology also resulted in either large cluster numbers too large to be adequately analysed manually, or with metrics not as optimal as those of the K-means solutions.

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑ Political	Education	*Prototypical cluster %* ↑ Political	Education
GWSD - Kmeans							
Cross Attention							
No dim. reduction	19	1.95	0.17	0.46	0.43	0.0	4.5
w/ PCA	17	**0.45**	**0.61**	0.53	0.53	0.0	0.0
w/ UMAP	18	1.05	0.49	0.51	0.44	22.4	0.0
Pooled Cross Attention							
No dim. reduction	16	2.76	0.07	0.43	0.44	5.7	0.0
w/ PCA	19	0.79	0.38	0.43	0.47	15.9	0.0
w/ UMAP	19	0.47	0.55	0.49	0.40	5.0	11.1
Encoder-Encoder							
No dim. reduction	18	5.77	0.02	0.53	0.34	27.6	**33.4**
w/ PCA	19	0.84	0.34	0.40	**0.60**	10.9	0.0
w/ UMAP	15	0.50	0.54	**0.69**	0.54	**40.3**	19.8
Classifier Model							
No dim. reduction	19	1.95	0.17	0.46	0.43	0.0	5.2
w/ PCA	17	0.45	**0.61**	0.53	0.53	0.0	0.0
w/ UMAP	18	1.05	0.49	0.51	0.44	21.8	0.0
Pretrained Decoder							
No dim. reduction	19	2.83	0.09	0.61	0.47	10.7	5.3
w/ PCA	19	0.47	0.59	0.42	0.44	16.1	0.0
w/ UMAP	17	0.52	0.53	0.51	0.58	0.0	0.0
Pretrained Encoder-Decoder							
No dim. reduction	19	2.53	0.06	0.48	0.55	4.9	4.9
w/ PCA	19	0.83	0.34	0.45	0.52	11.4	11.4
w/ UMAP	17	0.84	0.34	0.36	0.57	0.0	5.5

Table 8: Internal and external validation metrics for the K-means clustering technique on the GWSD dataset. Internal validation metrics explain intra-cluster separation through higher Silhouette and lower Davies-Bouldin (DB Index) scores. External validity, which indicates the potential of having captured a voice, is measured via the average Purity score and % of prototypical clusters.

C.6 Preliminary experiments with different model sizes

During our preliminary experiments, we ran the Encoder-Encoder model using both base and large variants for T5 v1.1. The hyperparameters used were the same between models and models were trained in an identical manner to control for possible confounding variables. We have included our preliminary results for both the supervised and unsupervised components of our model of the Encoder-Encoder architecture, in Table 13 and Table 14 respectively.

As shown by the results of the supervised component in Table 13, using the larger model led to marginal improvements in performance. The results of the unsupervised component can be found in Table 14. The base model largely follow the trends of the larger model shown in the paper, although the marginal differences in the supervised component cascade into larger differences in quality of prototypical clusters formed. This would indicate that models that are smaller in size might struggle to adequately capture annotating behaviours within the latent space.

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑ Political	Education	*Prototypical cluster %* ↑ Political	Education
GWSD -GMM							
Unpooled Cross Attention							
No dim. reduction	5	12.54	0.00	0.44	0.55	0.0	0.0
w/ PCA	5	8.13	0.00	0.44	0.55	0.0	0.0
w/ UMAP	5	8.02	0.01	0.44	0.55	0.0	0.0
Pooled Cross Attention							
No dim. reduction	6	3.73	0.04	0.46	**0.57**	0.0	0.0
w/ PCA	6	2.68	0.05	0.46	**0.57**	0.0	0.0
w/ UMAP	7	2.31	0.08	0.37	0.46	0.0	0.0
Encoder-Encoder							
No dim. reduction	5	9.30	0.01	0.44	0.47	0.0	0.0
w/ PCA	5	4.09	0.03	0.44	0.47	0.0	0.0
w/ UMAP	5	5.57	0.03	0.44	0.47	0.0	0.0
Classifier Model							
No dim. reduction	5	1.87	0.19	0.43	0.51	0.0	0.0
w/ PCA	5	1.48	**0.33**	0.43	0.51	0.0	0.0
w/ UMAP	12	3.02	0.05	0.42	0.50	8.3	0.0
Pretrained Decoder							
No dim. reduction	19	3.12	0.05	0.41	0.50	4.6	0.0
w/ PCA	6	1.72	0.18	0.44	0.48	0.0	0.0
w/ UMAP	5	**1.75**	0.20	**0.47**	0.53	0.0	0.0
Pretrained Encoder-Decoder							
No dim. reduction	5	3.39	0.05	**0.47**	0.48	0.0	0.0
w/ PCA	6	2.90	0.00	0.44	0.56	0.0	0.0
w/ UMAP	11	2.51	0.06	0.45	0.43	**8.8**	0.0

Table 9: Internal and external validation metrics for the GMM clustering technique on the GWSD dataset. Internal validation metrics explain intra-cluster separation through higher Silhouette and lower Davies-Bouldin (DB Index) scores. External validity, which indicates the potential of having captured a voice, is measured via the average Purity score and % of prototypical clusters.

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑ Political	Education	*Prototypical cluster %* ↑ Political	Education
GWSD- HDBSCAN							
Unpooled Cross Attention							
No dim. reduction	407	0.62	0.57	**1.00**	**1.00**	**95.9**	**100.0**
w/ PCA	4	10.10	0.05	0.50	0.50	24.6	50.2
w/ UMAP	3	17.35	0.01	0.57	0.57	33.0	33.0
Pooled Cross Attention							
No dim. reduction	191	1.25	0.30	0.80	0.60	58.8	35.1
w/ PCA	3	2.47	0.01	0.60	0.50	33.3	0.0
w/ UMAP	173	0.23	0.85	0.75	0.38	58.8	35.1
Encoder-Encoder							
No dim. reduction	4	9.53	0.01	0.67	0.67	50.0	24.9
w/ PCA	5	6.99	0.03	0.43	0.57	0.0	39.5
w/ UMAP	4	21.22	0.07	0.52	0.92	24.6	25.6
Classifier Model							
No dim. reduction	211	0.14	0.95	0.50	0.62	59.1	35.0
w/ PCA	210	**0.13**	0.95	0.50	0.62	58.8	34.3
w/ UMAP	3	3.20	0.14	0.51	0.42	0.0	0.0
Pretrained Decoder							
No dim. reduction	210	1.21	0.62	0.50	0.62	60.0	34.6
w/ PCA	204	1.14	0.52	0.40	0.60	56.5	38.2
w/ UMAP	210	0.78	**0.98**	0.50	0.62	58.9	34.7
Pretrained Encoder-Decoder							
No dim. reduction	3	0.72	0.25	0.50	0.50	33.0	33.0
w/ PCA	3	2.31	0.04	0.50	0.50	33.0	33.0
w/ UMAP	—	—	—	—	—	—	—

Table 10: Internal and external validation metrics for the HDBSCAN clustering technique on the GWSD dataset. Internal validation metrics explain intra-cluster separation through higher Silhouette and lower Davies-Bouldin (DB Index) scores. External validity, which indicates the potential of having captured a voice, is measured via the average Purity score and % of prototypical clusters. Missing runs indicate that the cluster number computed was equal to the amount of text-annotation pairs, proving the solution invalid.

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑		*Prototypical cluster %* ↑	
				Political	Education	Political	Education
MBIC- GMM							
Unpooled Cross Attention							
No dim. reduction	19	<u>7.50</u>	<u>0.01</u>	<u>**0.66**</u>	<u>0.54</u>	<u>**31.7**</u>	<u>5.2</u>
w/ PCA	5	8.11	0.00	0.41	0.46	0.0	0.0
w/ UMAP	5	8.22	0.00	0.41	0.46	0.0	0.0
Pooled Cross Attention							
No dim. reduction	19	<u>4.04</u>	<u>0.02</u>	0.37	0.46	<u>**32.2**</u>	<u>5.4</u>
w/ PCA	8	4.09	0.00	<u>0.45</u>	<u>**0.56**</u>	12.0	0.0
w/ UMAP	5	7.83	0.01	<u>0.45</u>	0.51	0.0	0.0
Encoder-Encoder							
No dim. reduction	19	<u>8.81</u>	0.00	<u>0.50</u>	0.33	<u>21.0</u>	<u>**21.4**</u>
w/ PCA	5	9.50	0.00	0.47	<u>0.48</u>	19.7	19.7
w/ UMAP	5	8.87	0.00	0.47	<u>0.48</u>	19.7	19.7
Classifier Model							
No dim. reduction	—	—	—	—	—	—	—
w/ PCA	—	—	—	—	—	—	—
w/ UMAP	—	—	—	—	—	—	—
Pretrained Decoder							
No dim. reduction	5	3.67	<u>0.03</u>	0.44	0.46	0.0	0.0
w/ PCA	16	<u>2.83</u>	0.01	0.52	0.32	0.0	0.0
w/ UMAP	18	7.50	0.01	<u>0.53</u>	<u>0.50</u>	<u>17.2</u>	0.0
Pretrained Encoder-Decoder							
No dim. reduction	6	1.76	0.14	0.47	0.46	0.0	0.0
w/ PCA	5	2.27	0.03	<u>0.49</u>	<u>0.48</u>	0.0	0.0
w/ UMAP	5	<u>**0.58**</u>	<u>**0.43**</u>	<u>0.49</u>	<u>0.48</u>	0.0	0.0

Table 11: Internal and external validation metrics for the GMM clustering technique on the GWSD dataset. Internal validation metrics explain intra-cluster separation through higher Silhouette and lower Davies-Bouldin (DB Index) scores. External validity, which indicates the potential of having captured a voice, is measured via the average Purity score and % of prototypical clusters. Rows with missing labels indicate inability of the GMM clustering technique to create a solution within the allotted train time for the respective configuration's hyperparameter sweep.

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑		*Prototypical cluster %* ↑	
				Political	Education	Political	Education
MBIC- HDBSCAN							
Unpooled Cross Attention							
No dim. reduction	862	1.01	0.71	**1.00**	**1.00**	**100.0**	**100.0**
w/ PCA	862	**0.86**	**0.72**	**1.00**	**1.00**	**100.0**	**100.0**
w/ UMAP	862	1.30	0.21	**1.00**	**1.00**	**100.0**	**100.0**
Pooled Cross Attention							
No dim. reduction	218	1.30	0.20	**1.00**	**1.00**	84.8	70.1
w/ PCA	218	1.26	0.29	**1.00**	**1.00**	84.8	70.1
w/ UMAP	218	2.85	0.80	**1.00**	**1.00**	84.8	70.1
Encoder-Encoder							
No dim. reduction	5	4.18	0.00	**1.00**	**1.00**	60.3	60.3
w/ PCA	3	3.59	0.06	**1.00**	**1.00**	67.0	33.2
w/ UMAP	5	4.10	0.04	**1.00**	**1.00**	60.4	59.9
Classifier Model							
No dim. reduction	3	2.70	0.15	0.50	0.58	0.0	0.0
w/ PCA	3	1.81	0.04	0.67	0.67	32.7	32.7
w/ UMAP	3	1.93	0.56	0.46	0.55	0.0	0.0
Pretrained Decoder							
No dim. reduction	185	1.22	0.45	0.44	0.56	68.3	44.2
w/ PCA	168	2.45	0.07	0.43	0.57	74.4	38.5
w/ UMAP	168	1.11	0.63	0.43	0.57	74.4	38.5
Pretrained Encoder-Decoder							
No dim. reduction	3	1.27	0.19	0.50	0.50	33.3	33.3
w/ PCA	3	2.78	0.04	0.53	0.47	33.3	0.0
w/ UMAP	3	3.29	0.08	0.53	0.49	33.3	0.0

Table 12: Internal and external validation metrics for the HDBSCAN clustering technique on the MBIC dataset. Internal validation metrics explain intra-cluster separation through higher Silhouette and lower Davies-Bouldin (DB Index) scores. External validity, which indicates the potential of having captured a voice, is measured via the average Purity score and % of prototypical clusters.

Model size	F1 Score ↑	Avg. Pairwise Similarity ↓
Base	0.70	0.22 ± 0.06
Large	0.72	0.21 ± 0.06

Table 13: Results of the supervised component with different model sizes using the pretrained T5 v1.1 model for the Encoder-Encoder architecture.

	# Clusters	DB Index ↓	Silhouette ↑	*Purity* ↑ Political	Education	*Prototypical cluster %* ↑ Political	Education
`google/t5-v1_1-base`							
No dim. reduction	19	7.21	0.01	0.51	0.51	**<u>47.4</u>**	**<u>26.3</u>**
PCA	19	**<u>0.49</u>**	**<u>0.54</u>**	0.50	**<u>0.52</u>**	5.3	5.2
UMAP	19	0.50	**<u>0.54</u>**	**<u>0.53</u>**	0.41	10.5	10.4
`google/t5-v1_1-large`							
No dim. reduction	19	6.93	0.01	0.41	0.46	21.1	15.8
PCA	19	**<u>0.49</u>**	**<u>0.54</u>**	**<u>0.53</u>**	0.43	15.0	7.7
UMAP	19	**<u>0.49</u>**	0.53	0.51	<u>0.48</u>	<u>36.8</u>	<u>21.1</u>

Table 14: Preliminary results of the unsupervised component for the Encoder-Encoder architecture using different sizes of the pretrained T5 v1.1 backbone.